Consider the coal data:

Measurement = % sulfur in plug taken from vein

Res Hyp: Sulfur content differs among the 5 veins under study

Minitab:

\[ KW = 19.08, \text{ DF } = 4, \text{ P } = 0.001 \]

Where is the significance?

Informally: look at 85% CI-Boxplots
Suggests: 1 and 3 differ from 4 and 5

Need formal statistical methods to confirm this.

Multiple comparisons
There are 5 veins.

There are $\frac{5 \times 4}{2} = 10$ pairwise comparisons.

Two sided Wilcoxon p-values

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.043</td>
<td>.340</td>
<td>.013</td>
<td>.0054</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>.229</td>
<td>.142</td>
<td>.076</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.003</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>.965</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Criteria for small p-value?
.05?
.01?

Why two sided?
How to decide?
Subhyp: ith vein is different from jth vein

\[ \frac{k \times (k-1)}{2} = 10 \text{ subhypotheses} \]

Suppose we reject a subhyp if p-value \( \leq 0.05 \)

Probability of a false rejection is .05

What is the probability of at least one false rejection among the 10 tests?

Cannot say exactly since the tests are interdependent.
Let FR$_i$ stand for false rejection by $ith$ test.

\[
P(\text{atleastonefalserejection}) =
\]

\[
P(FR_1 \text{ or } FR_2 \text{ or } \ldots \text{or } FR_{10}) \leq
\]

\[
P(FR_1) + P(FR_2) + \ldots + P(FR_{10}) = 10 \times .05 = .5
\]

So the probability of false rejection among the 10 tests is bounded above by $.5$ and may be close to $.5$.

So the individual criterion of $.05$ for each comparison leads to a large prob of at least one false rejection.
IDEA:

- You set the prob of at least one false rejection (say .1), called experiment-wise error rate (p92 text)
- Then budget the .1 over the 10 comparisons to get $0.1 / 10 = 0.01$ as criterion for individual p-values. This called comparison error rate.
- This guarantees the prob of at least one false rejection is no greater than .1.
For the coal veins:

1. Experiment-wise error rate .1
2. Comparison error rate \( \frac{1}{10} = .01 \)
3. This is the criterion for the 2-sided p-values

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.043</td>
<td>.340</td>
<td>.013</td>
<td>.0054</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.229</td>
<td>.142</td>
<td>.076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>.003</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>.965</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
When to use two-sided or one-sided multiple comparisons

If the res hyp does not specify any direction then use two-sided.

- If the res hyp specifies an expected pattern for experimental results then
- You should probably not use a general test like KW.
- You should tailor the test to the experiment.
- You should then only use the relevant one-sided comparisons.
- This will result in much more statistical power. But you must make these decisions before looking at the data.